Mid-term work----Jing Tang

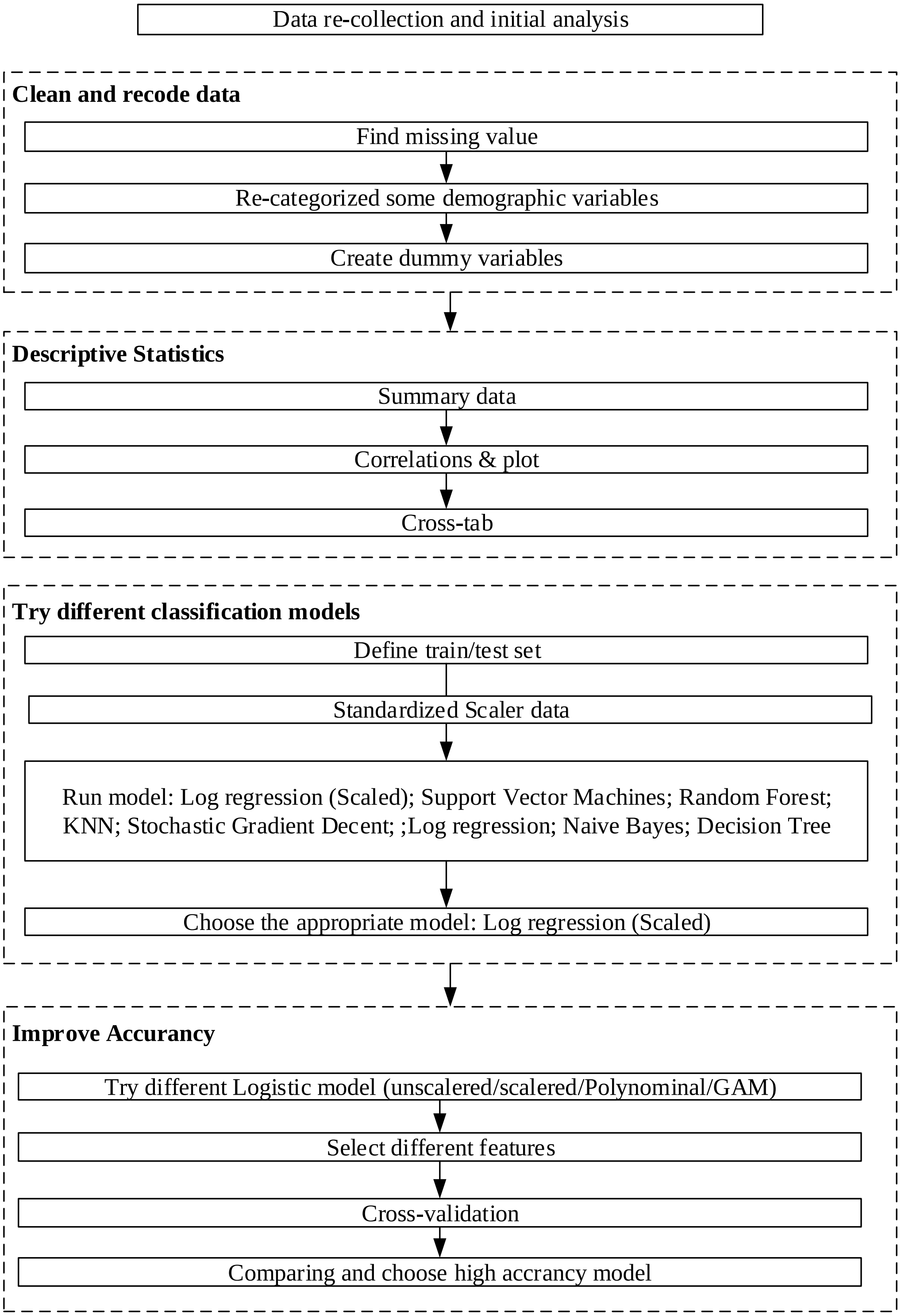
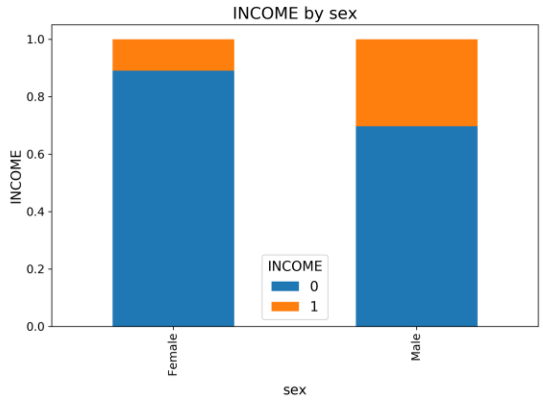
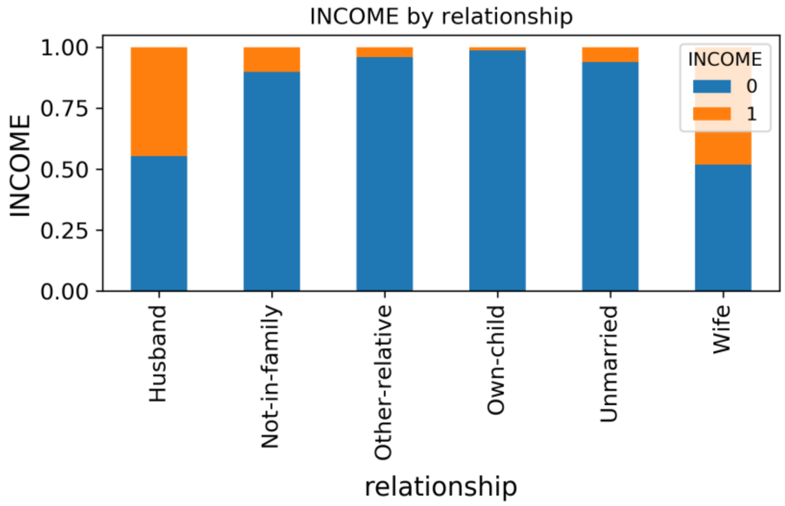
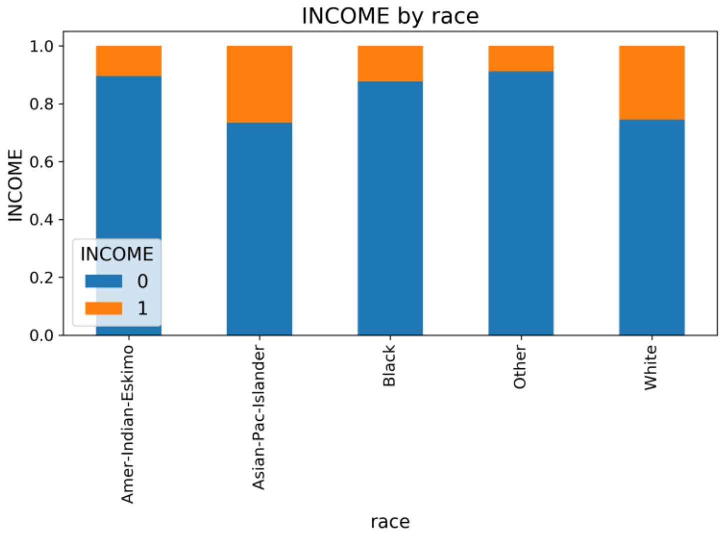

Data re-collection and initial analysis
Clean and recode data
Find missing value
Re-categorized some demographic variables
Create dummy variables
Descriptive Statistics
Summary data
Correlations & plot
Cross-tab
Try different classification models
Define train/test set
Standardized Scaler data
Run model: Log regression (Scaled); Support Vector Machines; Random Forest; KNN; Stochastic Gradient Decent; ;Log regression; Naive Bayes; Decision Tree
Choose the appropriate model: Log regression (Scaled)
Improve Accurancy
Try different Logistic model (unscalered/scalered/Polynominal/GAM)
Select different features
Cross-validation
Comparing and choose high accrancy model

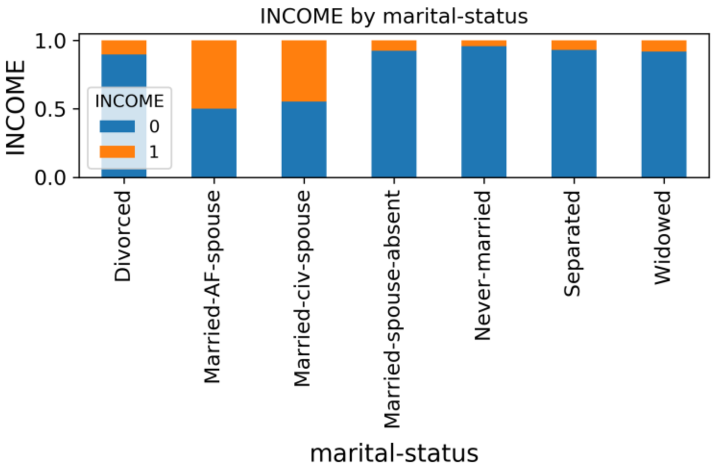
Figure 1 Roadmap Figure

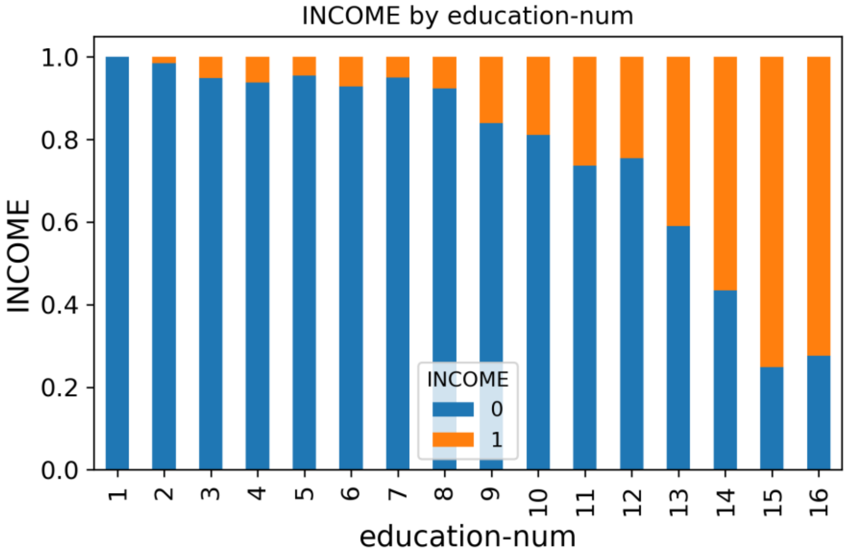
Exploratory analyses

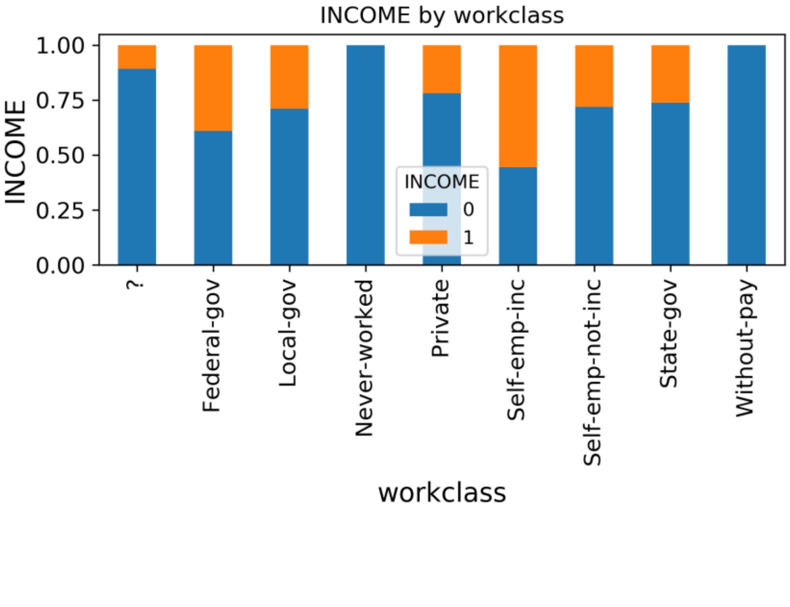


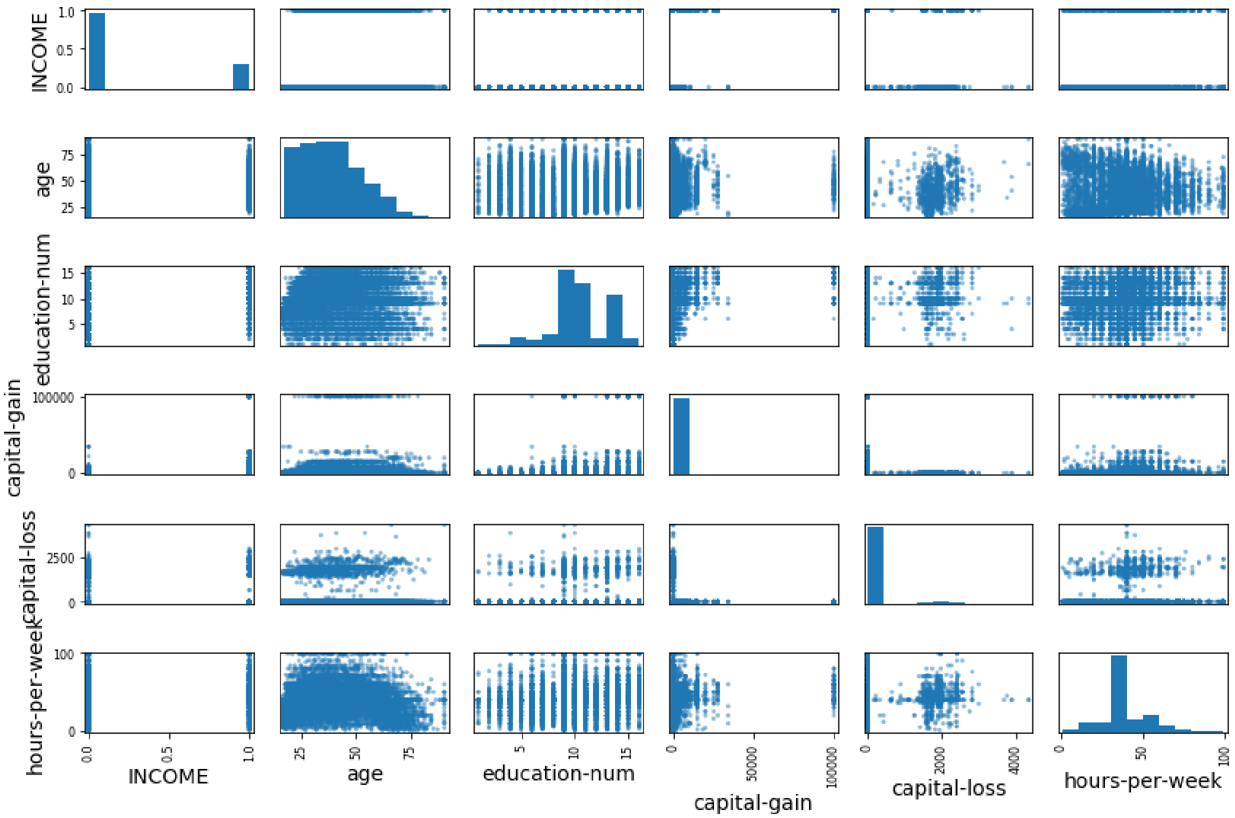


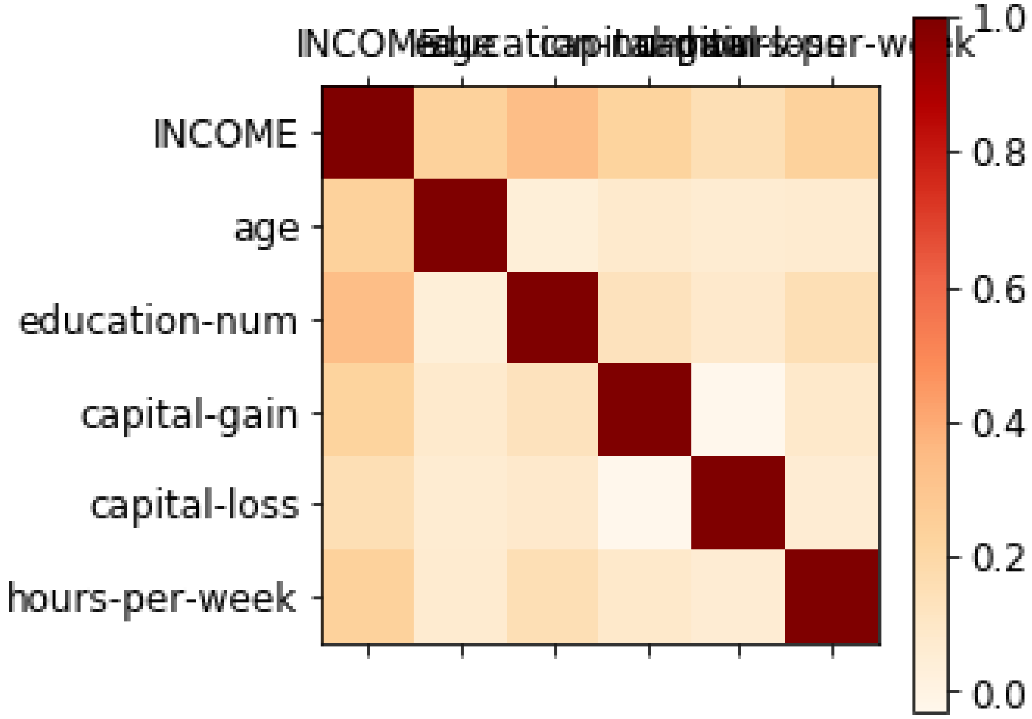



INCOME by marital-status
INCOME
1.0
0.5
0.0
INCOME
0
1
Divorced
Married-AF-spouse
Married-civ-spouse
Married-spouse-absent
Never-married
Separated
Widowed
marital-status


INCOME by education-num
INCOME
1.0
0.8
0.6
0.4
0.2
0.0
INCOME
0
1
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
education-num


INCOME by workclass
INCOME
1.00
0.75
0.50
0.25
0.00
INCOME
0
1
?
Federal-gov
Local-gov
Never-worked
Private
Self-emp-inc
Self-emp-not-inc
State-gov
Without-pay
workclass


INCOME
age
education-num
capital-gain
capital-loss
hours-per-week
1.0
0.5
0.0
25
50
75
5
10
15
100000
0
50000
2500
2000
4000
100


INCOME
age
education-num
capital-gain
capital-loss
hours-per-week
1.0
0.8
0.6
0.4
0.2
0.0

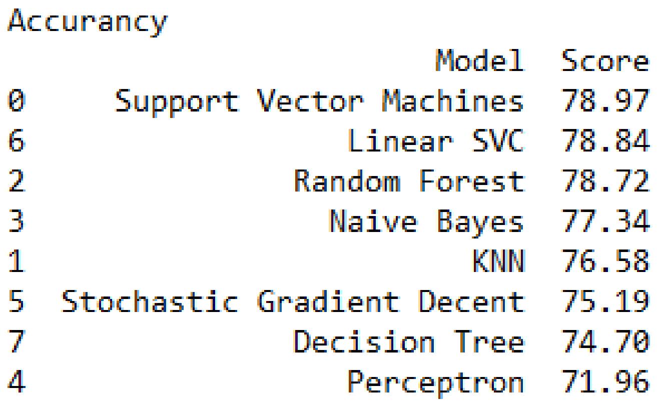
Summary

- Demographic variables matter.
- High correlated continuous variable: education-num, hours-per-week, capital-gain, age.
- Potential non-linear variables: hours-per-week, age

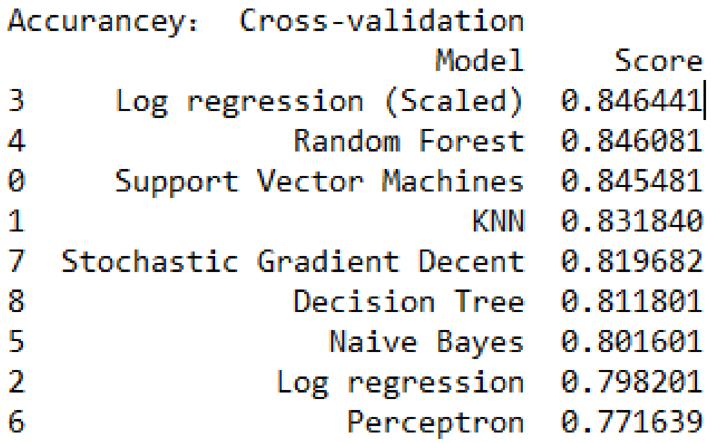
Building a prediction model

Try different model

```
feature_cols = ['age', 'fnlwgt', 'education-num', 'capital-gain', 'capital-loss',
        'hours-per-week', 'WORKC_1', 'WORKC_2', 'WORKC_3', 'WORKC_4', 'MARRI_1',
        'MARRI_2']
```

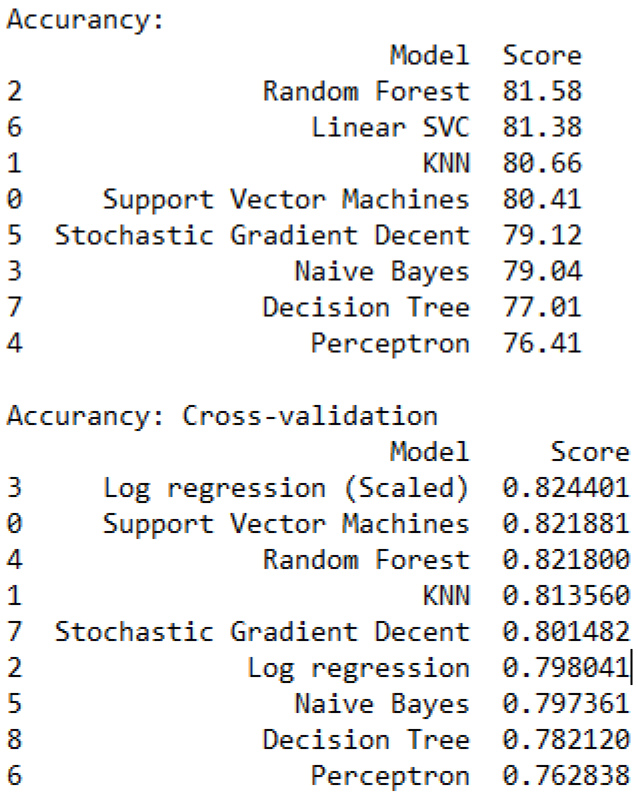


```
Accurancy
                        Model  Score
0     Support Vector Machines  78.97
6                  Linear SVC  78.84
2               Random Forest  78.72
3                 Naive Bayes  77.34
1                         KNN  76.58
5  Stochastic Gradient Decent  75.19
7               Decision Tree  74.70
4                  Perceptron  71.96
```



```
Accurancey:  Cross-validation
                        Model     Score
3     Log regression (Scaled)  0.846441
4               Random Forest  0.846081
0     Support Vector Machines  0.845481
1                         KNN  0.831840
7  Stochastic Gradient Decent  0.819682
8               Decision Tree  0.811801
5                 Naive Bayes  0.801601
2              Log regression  0.798201
6                  Perceptron  0.771639
```

feature_int = ['age', 'fnlwgt', 'education-num', 'capital-gain', 'capital-loss', 'hours-per-week', 'Gender']



```
Accurancy:
                          Model  Score
2                 Random Forest  81.58
6                    Linear SVC  81.38
1                           KNN  80.66
0       Support Vector Machines  80.41
5   Stochastic Gradient Decent  79.12
3                   Naive Bayes  79.04
7                 Decision Tree  77.01
4                    Perceptron  76.41

Accurancy: Cross-validation
                          Model     Score
3       Log regression (Scaled)  0.824401
0       Support Vector Machines  0.821881
4                 Random Forest  0.821800
1                           KNN  0.813560
7   Stochastic Gradient Decent  0.801482
2                Log regression  0.798041
5                   Naive Bayes  0.797361
8                 Decision Tree  0.782120
6                    Perceptron  0.762838
```

Summary:

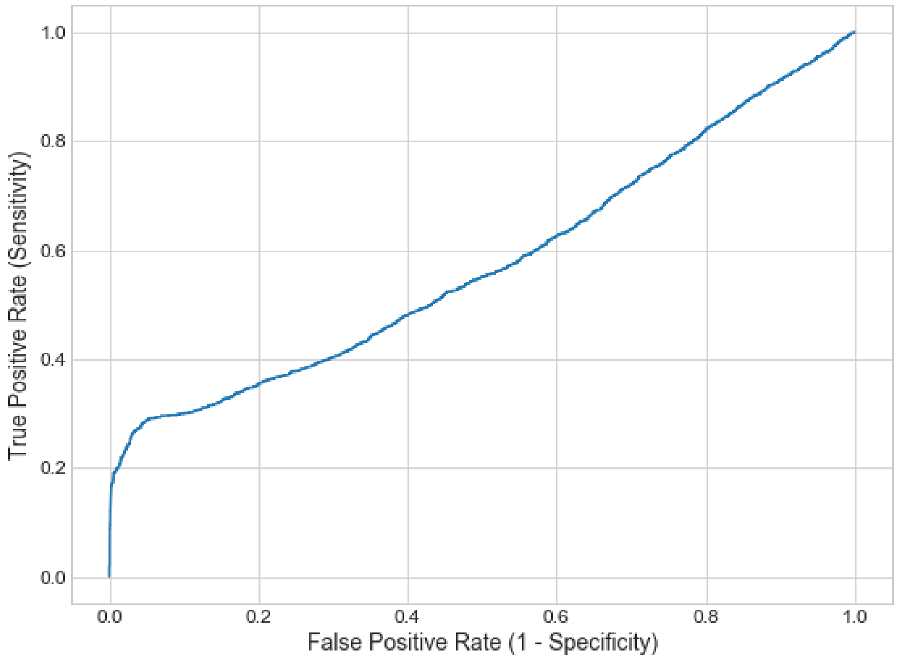
- Choosing Logistic model (Standardized data) as the original model, then improving the accuracy of Logistic model.

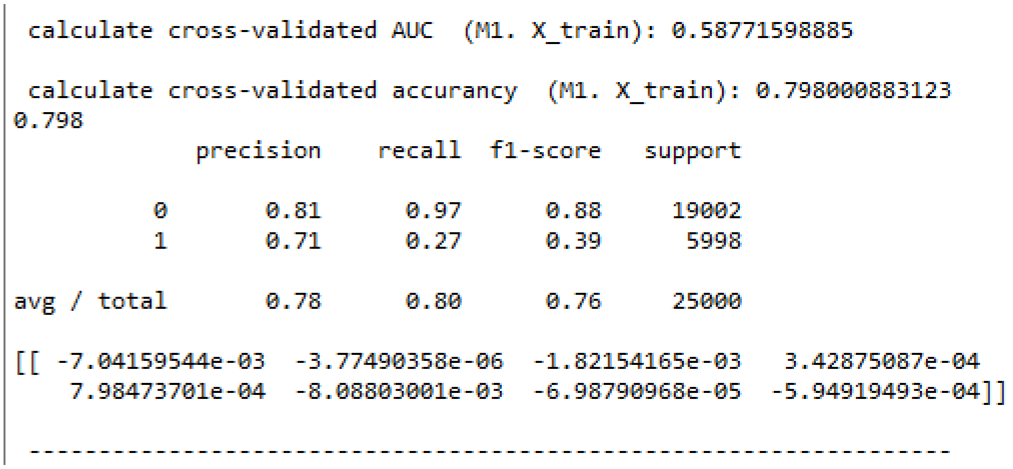
Logistic Model

Logistic Model0

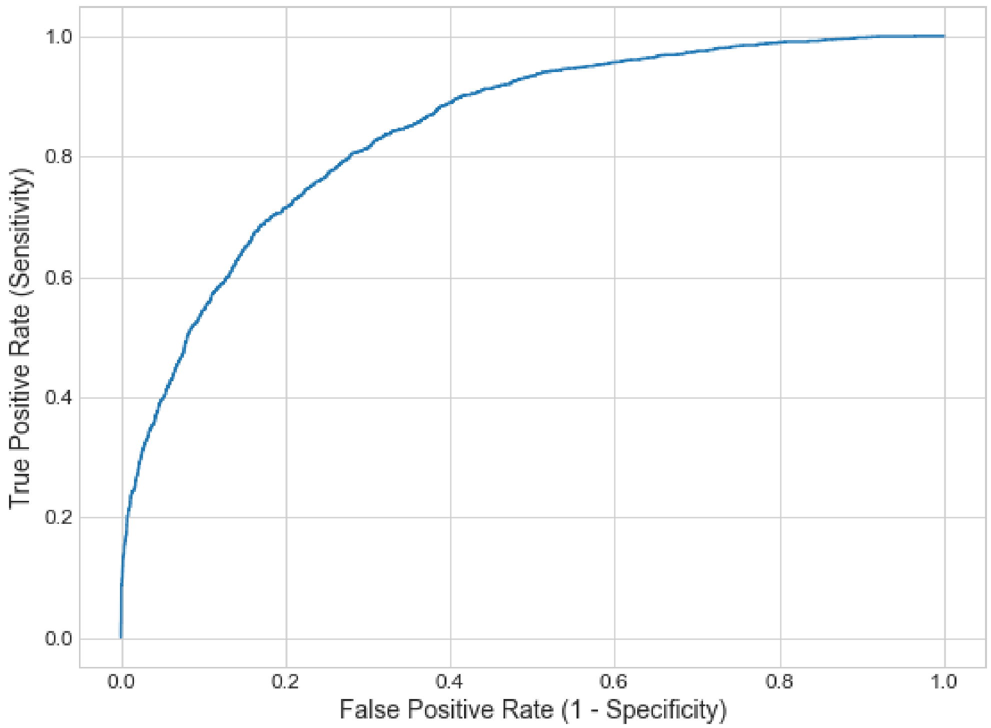
Unscaled Model

feature_int = ['age', 'fnlwgt', 'education-num', 'capital-gain', 'capital-loss', 'hours-per-week','COUNTRY','Gender']

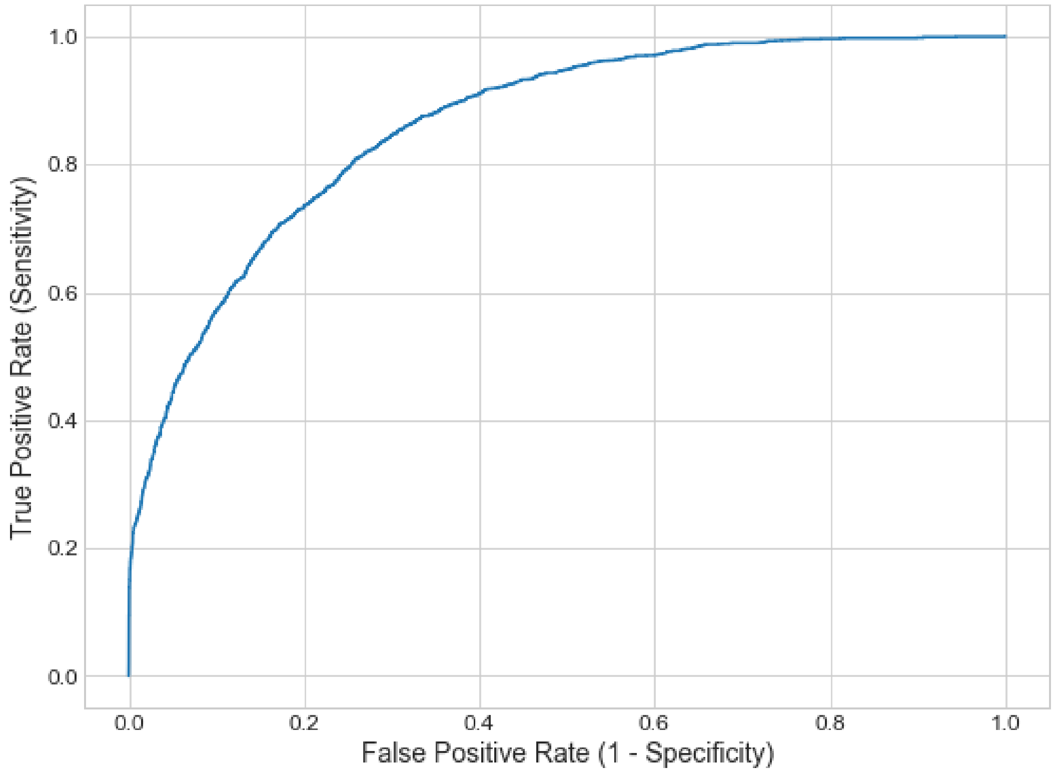




```
 calculate cross-validated AUC  (M1. X_train): 0.58771598885

 calculate cross-validated accurancy  (M1. X_train): 0.798000883123
0.798
             precision    recall  f1-score   support

          0       0.81      0.97      0.88     19002
          1       0.71      0.27      0.39      5998

avg / total       0.78      0.80      0.76     25000

[[ -7.04159544e-03  -3.77490358e-06  -1.82154165e-03   3.42875087e-04
    7.98473701e-04  -8.08803001e-03  -6.98790968e-05  -5.94919493e-04]]

 ------------------------------------------------------------------
```



```
 ------------------------------------------------------------------

 calculate cross-validated AUC  (M2. X_train_scaled): 0.847230754501

 calculate cross-validated accurancy  (M2. X_train_scaled): 0.825240820903
0.82524
             precision    recall  f1-score   support

          0       0.84      0.95      0.89     19002
          1       0.72      0.44      0.55      5998

avg / total       0.81      0.83      0.81     25000

[[ 0.58573414  0.05504313  0.86491518  2.36424125  0.28049507  0.43726323
  -0.06905303 -0.54412693]]

 ------------------------------------------------------------------
```



Area under the ROC curve.: 0.861541415606

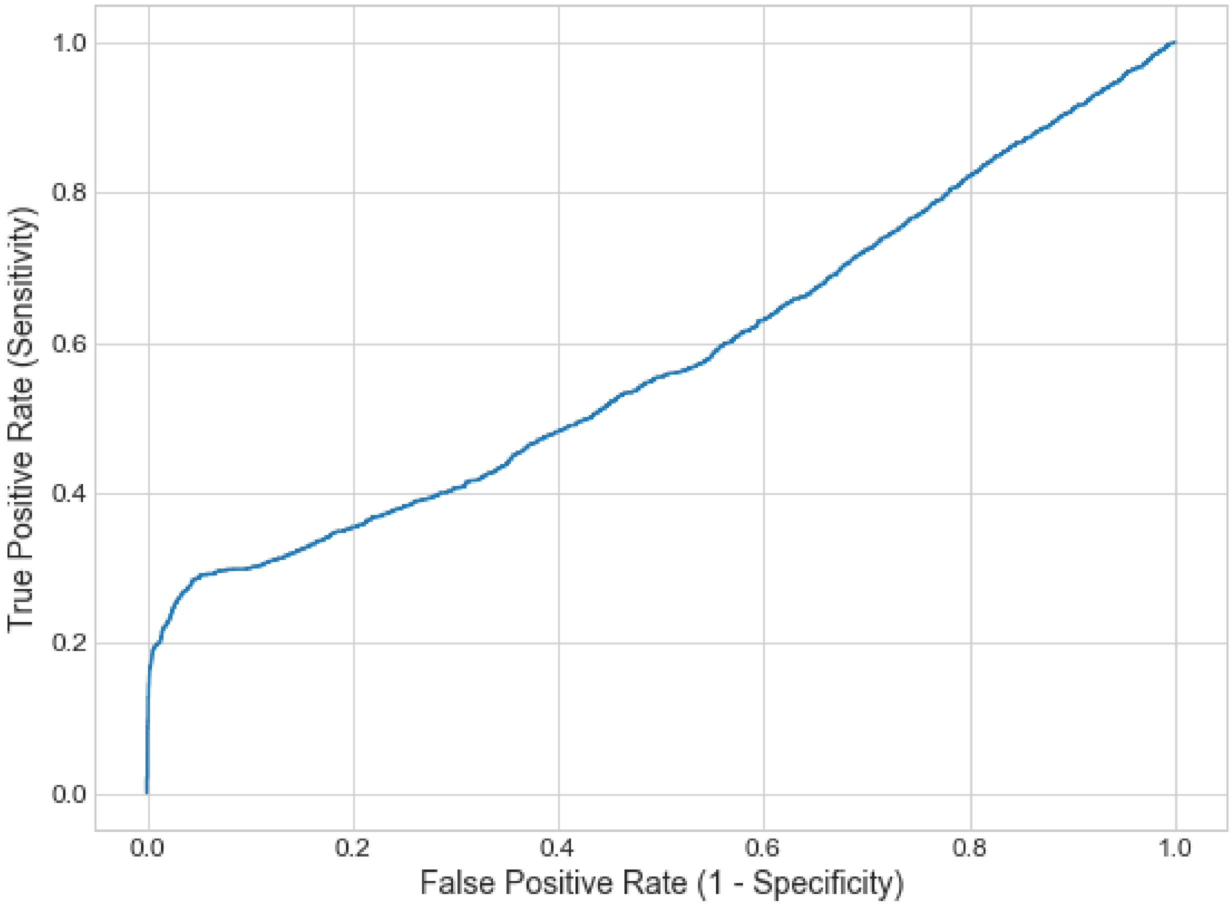
calculate testing accuracy (M1. X_train_scaled_poly): 0.825552175638

Logistic Model1 (Full model)

Full model: feature_cols = ['age', 'fnlwgt', 'education-num', 'capital-gain', 'capital-loss','RACE_1','RACE_2','RACE_3','RACE_4',   'hours-per-week',   'WORKC_1', 'WORKC_2' ,'WORKC_3', 'WORKC_4' ,'MARRI_1', 'MARRI_2', 'COUNTRY','RELATION_1', 'RELATION_2', 'RELATION_3', 'RELATION_4', 'RELATION_5', 'OCCUP_1',

'OCCUP_2', 'OCCUP_3' ,'OCCUP_4', 'OCCUP_5' ,'OCCUP_6', 'OCCUP_7', 'OCCUP_8',   'OCCUP_9', 'OCCUP_10', 'OCCUP_11', 'OCCUP_12', 'OCCUP_13', 'OCCUP_14']

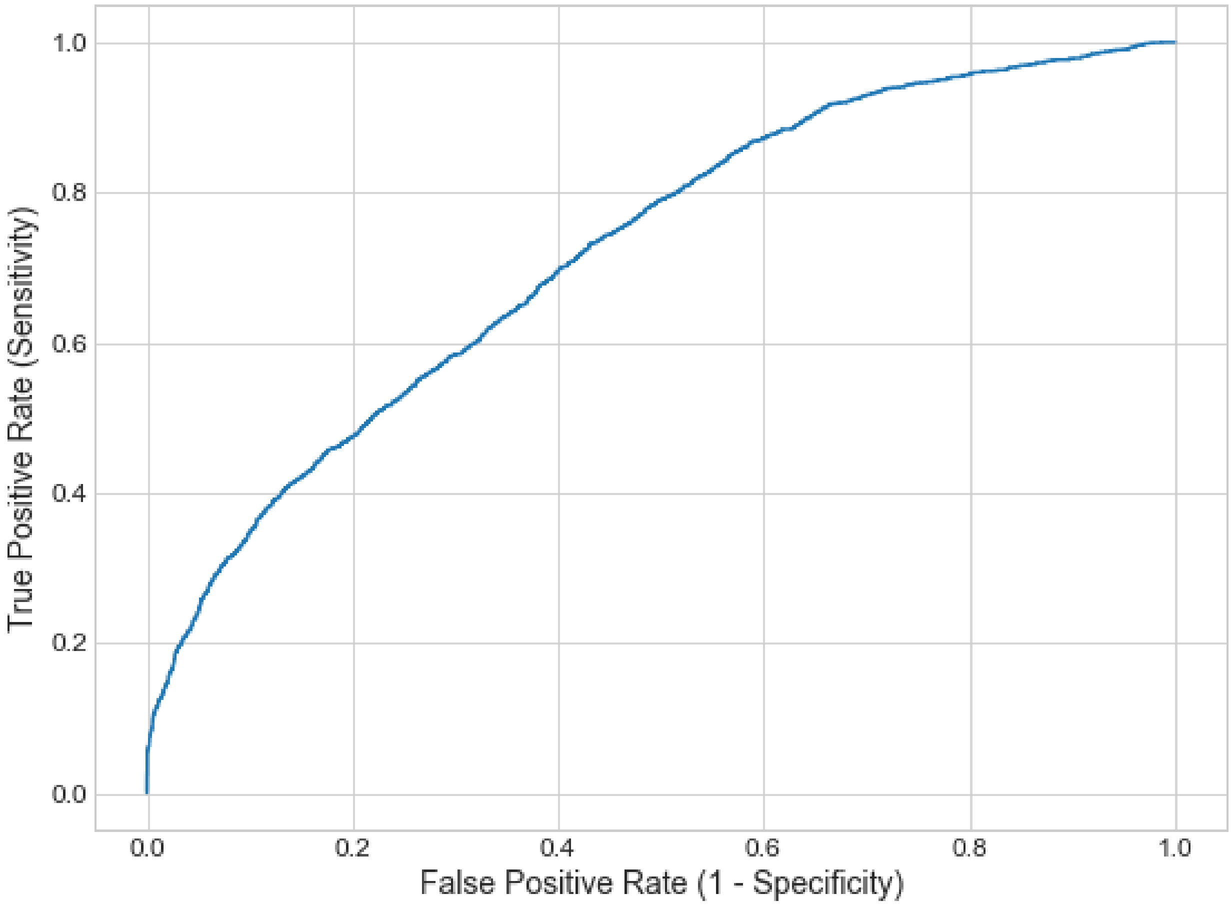
Unscaled Model



calculate testing accuracy (M1. X_train): 0.795265176564

Area under the ROC curve.: 0.578371088093

Standardized-scaled model



calculate testing accuracy (M1. X_train_scaled): 0.758497553234

Area under the ROC curve.: 0.720032426562

Estimates:

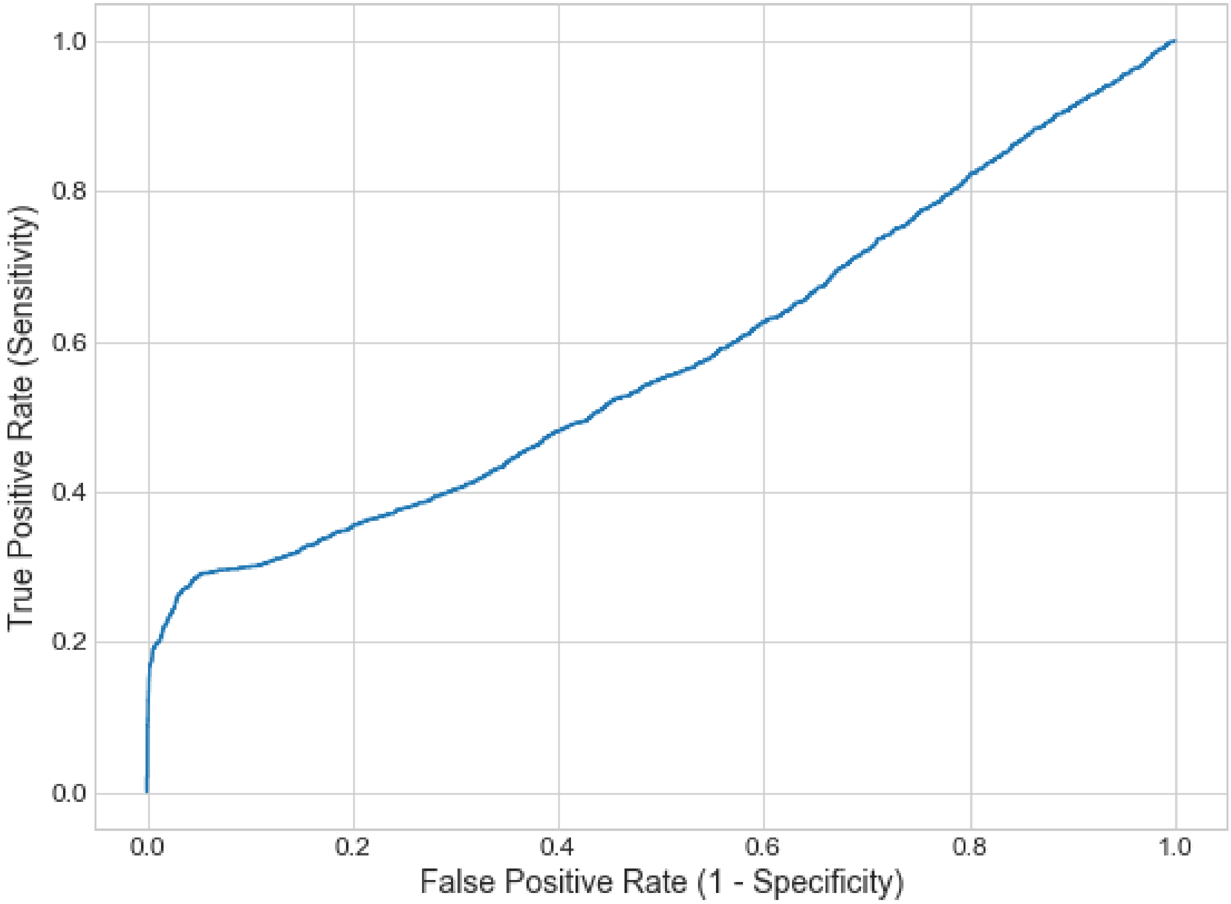
```
['age',  'fnlwgt',  'education-num', 'capital-gain', 'capital-loss',
'RACE_1','RACE_2','RACE_3','RACE_4' 'hours-per-week',   'WORKC_1',
'WORKC_2' ,'WORKC_3', 'WORKC_4' ,'MARRI_1', 'MARRI_2',
'COUNTRY','RELATION_1', 'RELATION_2', 'RELATION_3', 'RELATION_4',
'RELATION_5', 'OCCUP_1', 'OCCUP_2', 'OCCUP_3' ,'OCCUP_4', 'OCCUP_5' ,'OCCUP_6',
'OCCUP_7', 'OCCUP_8',      'OCCUP_9', 'OCCUP_10', 'OCCUP_11', 'OCCUP_12',
'OCCUP_13', 'OCCUP_14']
```

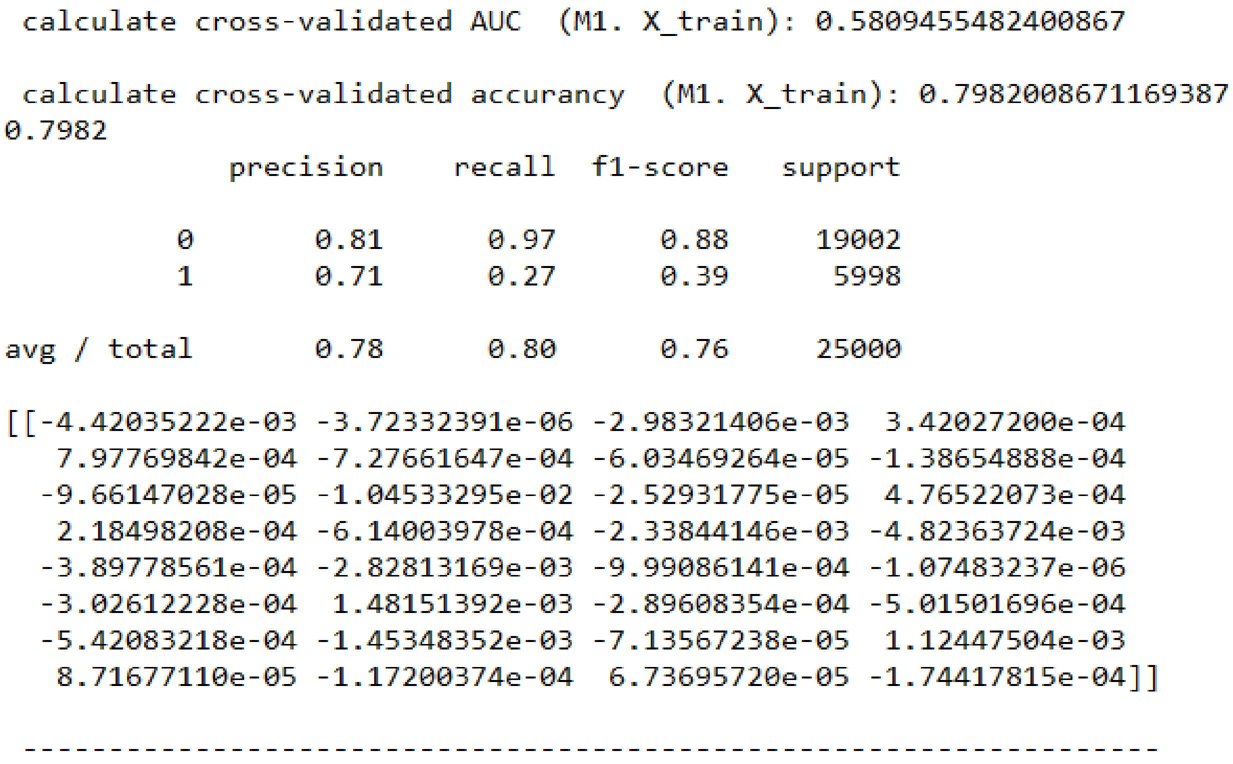
```
[[ 0.3749525   0.07951203  0.72563887  2.39787933  0.26457073 -0.04061072

   0.01748757 -0.05566635 -0.05892348  0.40488188 -0.12655655  0.02862326

  -0.03228724 -0.94654044 -0.96323491 -1.31738909 -0.0828622   0.15580986

  -0.08359726 -0.32238962 -0.0126279   0.11661218 -1.10978715 -0.0564222
  -1.11278668 -0.88370349 -0.74886455 -0.79963161 -0.91424917 -1.32847974
  -0.49463583 -0.98504778 -0.40401746 -1.00351381 -0.455645   -0.75747635]]
```

Logistic Model2 (remove ‘relationship’)

```
feature_cols1 = ['age',  'fnlwgt',  'education-num', 'capital-gain', 'capital-
loss','RACE_1','RACE_2','RACE_3','RACE_4', 'hours-per-week',   'WORKC_1',
'WORKC_2' ,'WORKC_3', 'WORKC_4' ,'MARRI_1', 'MARRI_2',    'COUNTRY', 'OCCUP_1',
'OCCUP_2', 'OCCUP_3' ,'OCCUP_4', 'OCCUP_5' ,'OCCUP_6', 'OCCUP_7', 'OCCUP_8',
'OCCUP_9', 'OCCUP_10', 'OCCUP_11', 'OCCUP_12', 'OCCUP_13', 'OCCUP_14']
```

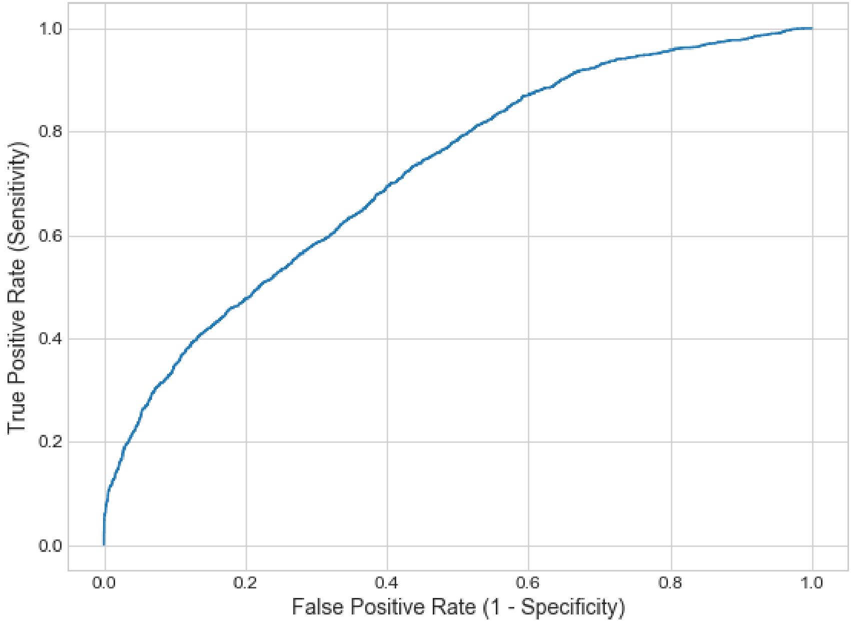
Unscaled Model

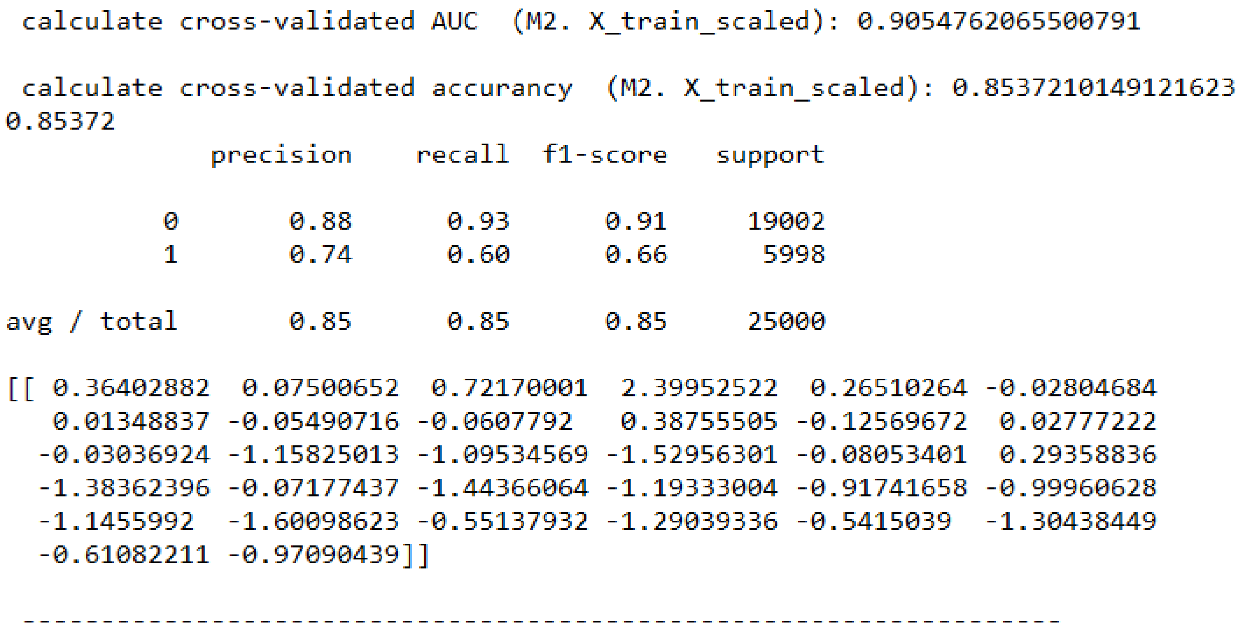




```
 calculate cross-validated AUC  (M1. X_train): 0.5809455482400867

 calculate cross-validated accurancy  (M1. X_train): 0.7982008671169387
0.7982
             precision    recall  f1-score   support

          0       0.81      0.97      0.88     19002
          1       0.71      0.27      0.39      5998

avg / total       0.78      0.80      0.76     25000

[[-4.42035222e-03 -3.72332391e-06 -2.98321406e-03  3.42027200e-04
   7.97769842e-04 -7.27661647e-04 -6.03469264e-05 -1.38654888e-04
  -9.66147028e-05 -1.04533295e-02 -2.52931775e-05  4.76522073e-04
   2.18498208e-04 -6.14003978e-04 -2.33844146e-03 -4.82363724e-03
  -3.89778561e-04 -2.82813169e-03 -9.99086141e-04 -1.07483237e-06
  -3.02612228e-04  1.48151392e-03 -2.89608354e-04 -5.01501696e-04
  -5.42083218e-04 -1.45348352e-03 -7.13567238e-05  1.12447504e-03
   8.71677110e-05 -1.17200374e-04  6.73695720e-05 -1.74417815e-04]]

 ------------------------------------------------------------------
```

Standardized-scaled model





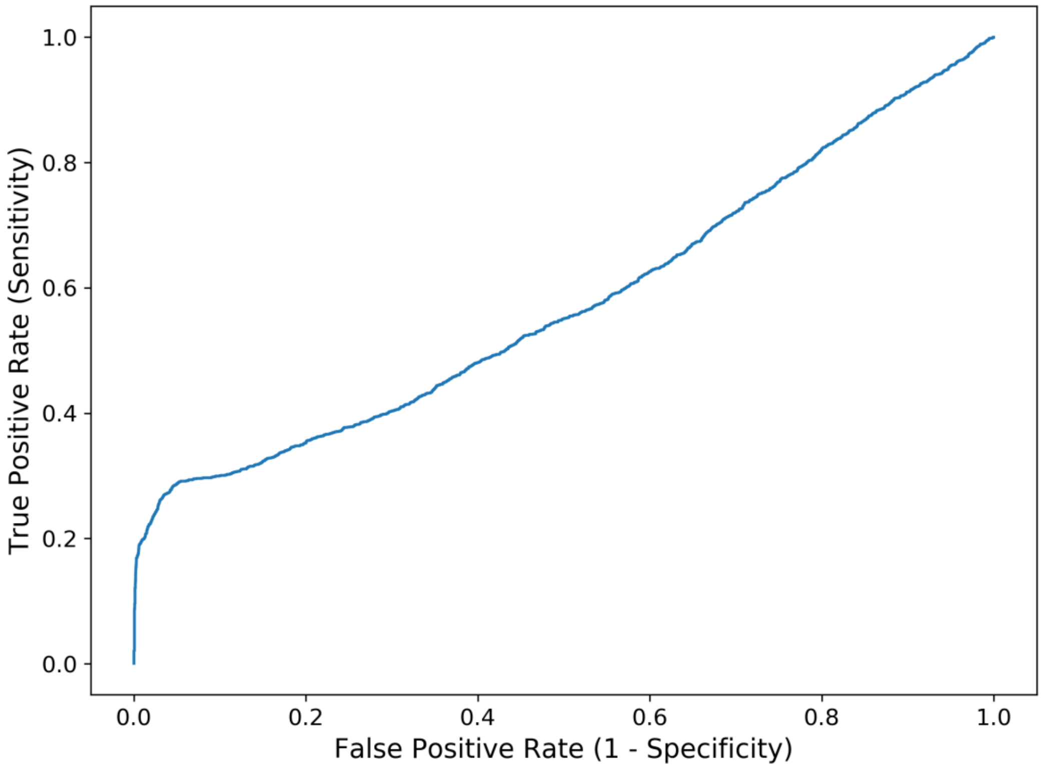
```
 calculate cross-validated AUC  (M2. X_train_scaled): 0.9054762065500791

 calculate cross-validated accurancy  (M2. X_train_scaled): 0.8537210149121623
0.85372
             precision    recall  f1-score   support

          0       0.88      0.93      0.91     19002
          1       0.74      0.60      0.66      5998

avg / total       0.85      0.85      0.85     25000

[[ 0.36402882  0.07500652  0.72170001  2.39952522  0.26510264 -0.02804684
   0.01348837 -0.05490716 -0.0607792   0.38755505 -0.12569672  0.02777222
  -0.03036924 -1.15825013 -1.09534569 -1.52956301 -0.08053401  0.29358836
  -1.38362396 -0.07177437 -1.44366064 -1.19333004 -0.91741658 -0.99960628
  -1.1455992  -1.60098623 -0.55137932 -1.29039336 -0.5415039  -1.30438449
  -0.61082211 -0.97090439]]

 ------------------------------------------------------------------
```

Standardized-scaled polynomial model (degree = 3)

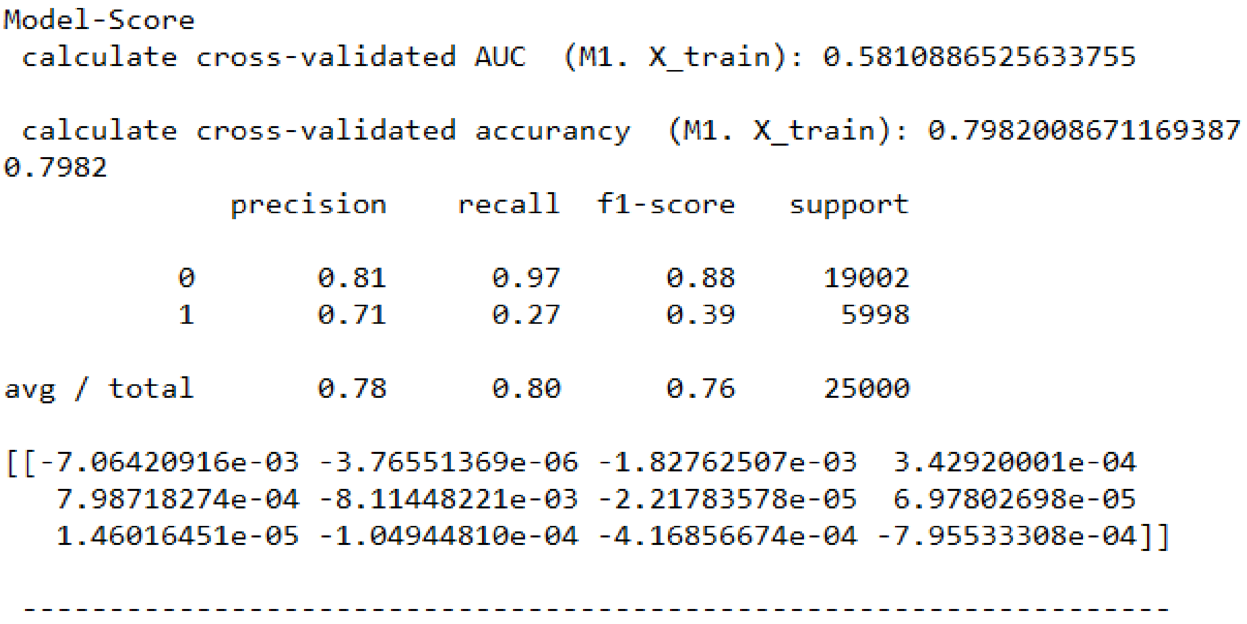
Logistic Model3

feature_cols = ['age', 'education-num','fnlwgt', 'capital-gain', 'capital-loss', 'hours-per-week', 'Gender', 'RACE_1','RACE_2','RACE_3','RACE_4','WORKC_1','WORKC_2','WORKC_3','WORKC_4','MARRI_1','MARRI_2']

Unscaled model

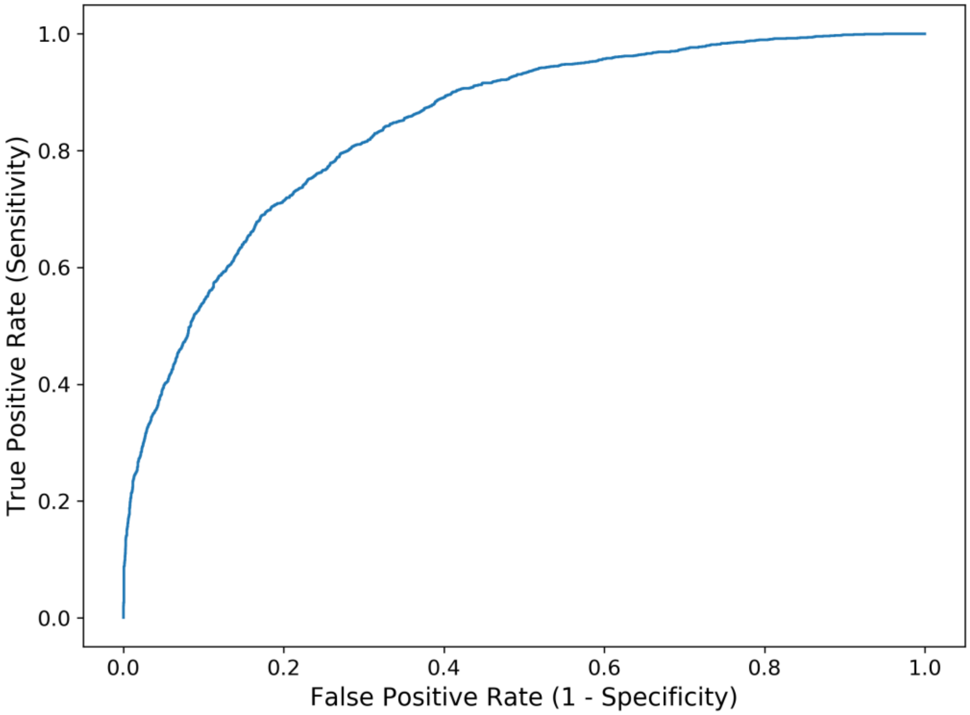


Mid.2.ROC curve (M1. X_train)

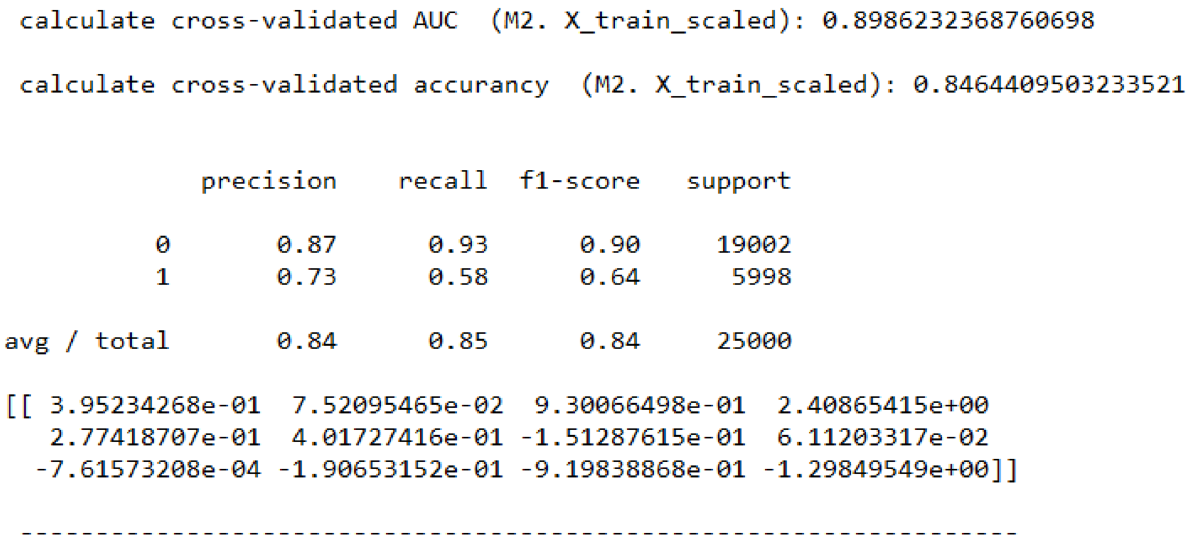


```
Model-Score
 calculate cross-validated AUC  (M1. X_train): 0.5810886525633755

 calculate cross-validated accurancy  (M1. X_train): 0.7982008671169387
0.7982
             precision    recall  f1-score   support

          0       0.81      0.97      0.88     19002
          1       0.71      0.27      0.39      5998

avg / total       0.78      0.80      0.76     25000

[[-7.06420916e-03 -3.76551369e-06 -1.82762507e-03  3.42920001e-04
   7.98718274e-04 -8.11448221e-03 -2.21783578e-05  6.97802698e-05
   1.46016451e-05 -1.04944810e-04 -4.16856674e-04 -7.95533308e-04]]

 -----------------------------------------------------------------
```

Standardized-scaled model

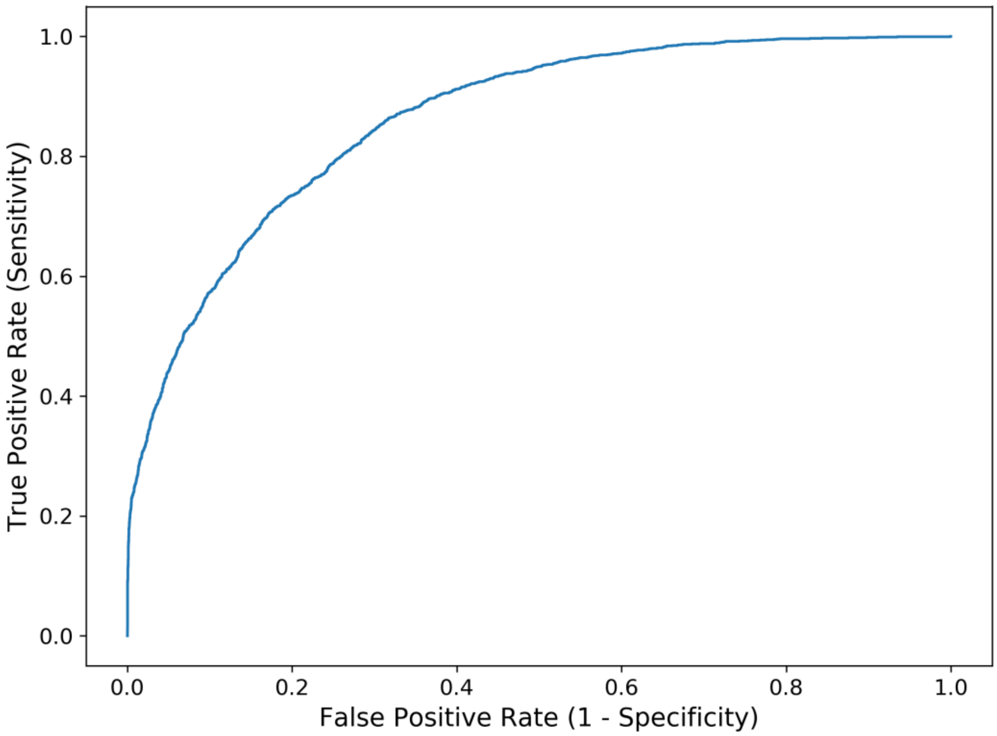


Mid.2.ROC curve (M1. X_train_scaled)

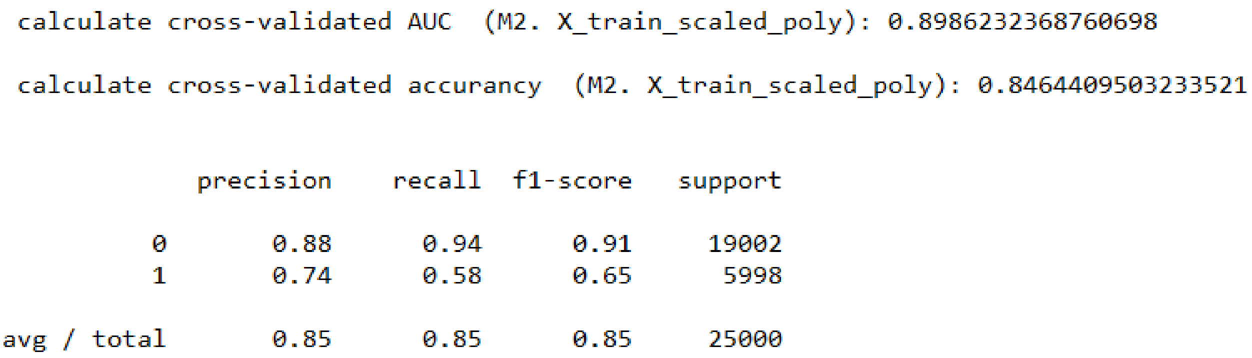


```
 calculate cross-validated AUC  (M2. X_train_scaled): 0.8986232368760698

 calculate cross-validated accurancy  (M2. X_train_scaled): 0.8464409503233521


             precision    recall  f1-score   support

          0       0.87      0.93      0.90     19002
          1       0.73      0.58      0.64      5998

avg / total       0.84      0.85      0.84     25000

[[ 3.95234268e-01  7.52095465e-02  9.30066498e-01  2.40865415e+00
   2.77418707e-01  4.01727416e-01 -1.51287615e-01  6.11203317e-02
  -7.61573208e-04 -1.90653152e-01 -9.19838868e-01 -1.29849549e+00]]

 ------------------------------------------------------------------
```

Standardized-scaled polynomial model (degree = 3)



Mid.2.ROC curve (M1. X_train_scaled_poly)



```
 calculate cross-validated AUC  (M2. X_train_scaled_poly): 0.8986232368760698

 calculate cross-validated accurancy  (M2. X_train_scaled_poly): 0.8464409503233521


             precision    recall  f1-score   support

          0       0.88      0.94      0.91     19002
          1       0.74      0.58      0.65      5998

avg / total       0.85      0.85      0.85     25000
```

Summary:

- Comparing to Model0 and Model1, Model2 and Model3 are better model.
- There are no need to include all the predictors in the model.
- Standardized-scaled polynomial model always show best performance.